



Representing Numerical Data Floating Point

**Reading: Chapter 5.3-5.4
(except details of floating point operations)**



Learning Objectives

- Understand the components of a floating point number
- Understand of floating point range of values and precision
- Understand IEEE floating point spec



Exponential Notation

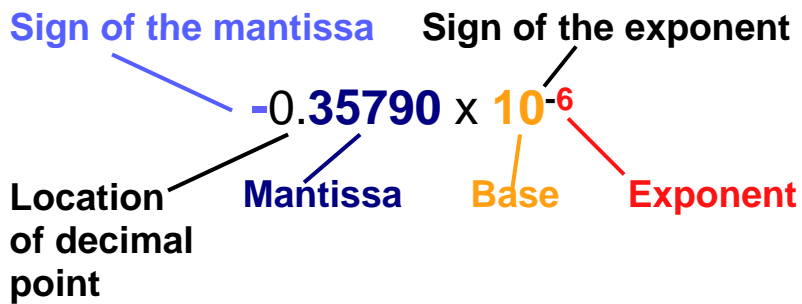
- Also called *scientific notation*
- Many ways to express a number – e.g., 12345 is
 - 12345
 - 12345×10^0
 - 0.12345×10^5
 - 123450000×10^{-4}
- Explicit specifications required for a number
 1. Sign (e.g., “+”) Mantissa is referred to as a significand in the IEEE spec
 2. Magnitude or mantissa (e.g., 12345)
 3. Sign of the exponent (e.g., “+”)
 4. Magnitude of the exponent (e.g., 5) Radix point is like a decimal point, but not restricted to base 10 (decimal) notation
- Plus (implied specs)
 1. Base of the exponent (10)
 2. Location of decimal point (or other base) radix point

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

5-3



Summary of Rules



Sign of the exponent usually represented with a bias. Bias is the number added to the exponent to get the stored exponent (e.g., IEEE spec states a bias of 127 for negative exponents)

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

5-4



Format Specification

- Internal format and interchange format
- Predefined format, usually multiples of 16 bits (e.g., 32 and 64)

Sign of the mantissa

SEEMMMM

Exponent

Mantissa

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

5-5



Radix Point

- For base 2, the usual term is binary point
- Assume radix point located at beginning of mantissa
- For example 5.0_{10} represented as $.101 \times 2^3$

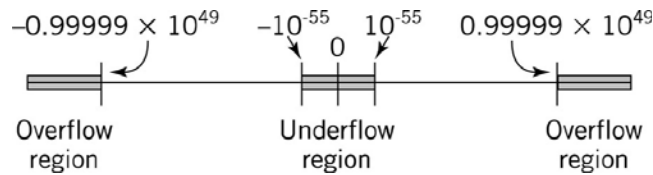
Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

5-6



Overflow and Underflow

- Possible for the number to be too large or too small for representation



Using infinity as a value allows operations to continue past overflow



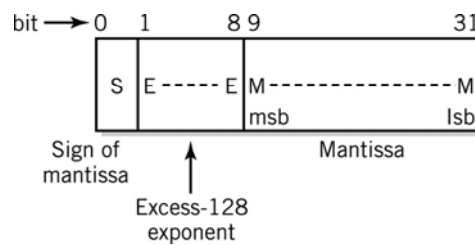
Floating Point Calculations

- Addition and subtraction
 - Exponent and mantissa treated separately
 - Exponents of numbers must agree
 - Align decimal points
 - Least significant digits may be lost
 - Mantissa overflow requires exponent again shifted right



Floating Point in the Computer

- Typical floating point format
 - 32 bits provide range $\sim 10^{-38}$ to 10^{+38}
 - 8-bit exponent = 256 levels
 - Excess-128 notation
 - 23/24 bits of mantissa: approximately 7 decimal digits of precision



Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

5-9



IEEE 754 Floating Point Standard

- Standard defines
 - Arithmetic formats
 - Interchange formats
 - Rounding algorithms
 - Operations
 - Exception handling
 - Extension recommendations
- Standard includes binary and decimal components
- Operations with infinite values are well defined

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

5-10



Programming Considerations

- Integer advantages
 - Easier for computer to perform
 - Potential for higher precision
 - Faster to execute
 - Fewer storage locations to save time and space

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

5-11



Programming Considerations

- Real numbers
 - Variable or constant has fractional part
 - Numbers take on very large or very small values outside integer range
 - Program should use least precision sufficient for the task

Copyright 2010-2011 John Wiley & Sons, Inc. & Robert F. Kelly

5-12